Published in final edited form as:

Comput Inform Nurs. 2017 May ; 35(5): 221–225. doi:10.1097/CIN.0000000000000359.

The Inherent Challenges of Using Large Datasets in Healthcare Research: Experiences of an Interdisciplinary Team

Aaron Kaulfus, M.S.,

Institution: University of Alabama in Huntsville, Department: Atmospheric Science Department,
Address: National Space Science and Technology Center, 320 Sparkman Dr., Huntsville, AL
35805

Susan Alexander, DNP, CRNP, ADM-BC,

Institution: University of Alabama in Huntsville, Department: College of Nursing

Shuang Zhao, Ph.D.,

Institution: University of Alabama in Huntsville, Department: Political Science and Atmospheric
Science Departments

Robert A. Oster, Ph.D.,

Institution: University of Alabama at Birmingham, Department: Medicine, Division: Preventive
Medicine

Louise O’Keefe, PhD, CRNP, and

Institution: University of Alabama in Huntsville, Department: College of Nursing

Al Bartolucci, Ph.D. [Professor, Emeritus]

Institution: University of Alabama at Birmingham, Department: Department of Biostatistics

Introduction

The use of big data has established a place across academia and business to enhance decision-making and inquiry into large-scale problems, tasks which were difficult in the past due to the limitations of small datasets in making connections across subjects and identifying broader patterns.^{1,2,3,4} While no formal definition of big data has been established, a review of research conducted using large datasets reveals three defining features: volume, variety, and velocity.^{2,5} In healthcare, large datasets (which may be a separate entity or extracted from big data) have the capacity to integrate activities of clinicians, administrators, policy makers, patients and researchers, possessing tremendous value for healthcare researchers.^{3,6,7} Queried correctly, datasets can be used to answer a variety of important questions at both regional and national levels, relating to hospital stays, emergency department visits, costs, quality of care, and others. The purpose of this paper is to discuss the experiences of using a large dataset to conduct cross-disciplinary healthcare research involving nursing, atmospheric science, and political science. Inherent challenges in using the large datasets quickly emerged, initially related to the manipulation and analysis of

the dataset, and followed by interpretation of findings within the inter-disciplinary and multi-institutional team.

The purpose of our team's research project was to identify and analyze relationships between inpatient hospital admissions for specified diagnoses and fluctuations in air quality. In addition to environmental data, the team used discharge data from the Nationwide Inpatient Sample (NIS), Healthcare Cost and Utilization Project (HCUP), Agency for Healthcare Research and Quality, years 2007 to 2011.^{8,9} The HCUP databases are a collection of datasets containing data collected from many states on topics such as patient encounters in emergency departments and inpatient hospital admissions to ambulatory surgeries within the United States. New databases are created annually and maintained by the Agency for Healthcare Research and Quality (AHRQ) through a federal-state-industry partnership. The NIS database, used for the study, contains data from more than 7 million hospital stays annually.^{8,9} While the HCUP NIS dataset was a valuable resource for addressing research questions, our research team faced a great deal of planning, along with unanticipated challenges, in extracting and analyzing the data.

Healthcare Quality and Utilization Project Datasets

With more than 100 variables currently existing in the HCUP NIS datasets, healthcare researchers have developed a mechanism to investigate a seemingly limitless variety of research questions with these robust secondary data. Data have been collected since 1988, and include contributions from 44 states (Figure 1). Participating hospitals are assigned identification numbers using state Federal Information Process Standards (FIPS) codes and zip codes. Combining the NIS dataset with the Nationwide Emergency Department Sample (NEDS) and Kids' Inpatient Dataset (KID), approximately 40 million patient entries are available from the AHRQ annually from 2007 to 2011. The total number of entries may change dramatically with state participation. Costs of purchasing the datasets are relatively inexpensive, ranging from \$350–\$500 depending on the year of availability. Additional supplementary data files, typically not available when the HCUP databases are originally released but designed to be linked to the purchased database, are also available free of charge. Tools needed to successfully use the datasets include a DVD drive, minimum 15 gigabytes of space available on a hard drive for each year of data, a third-party file compression utility, and statistical analysis software. To assist researchers, AHRQ offers load programs for each statistical analysis software package on its Web site.^{8,9} None of these software packages are open-source, which may limit interested users, however community users have developed some programming tools, including those for the commonly used Python language (PyHCUP)¹⁶, for working with the HCUP data. In addition, data format specification files are provided by AHRQ such that a researcher is able to develop load and data manipulation code using a preferred analysis software.

Use of Healthcare Cost and Utilization Project data in the literature

The HCUP databases have been widely used by public and private sector researchers. A search of the PubMed database using the general search term of 'HCUP database' yielded 511 publications, in peer-reviewed journals, with dates ranging from 1996 to 2016. In the

cited publications, HCUP databases were used for analysis of topics such as emerging treatments, surgery rates and surgical complications, predictors of outcomes, and population-based mortality risk factors. In addition to use by private researchers, a search of the HCUP Web site reveals a publication search option used to find technical reports, incorporating HCUP data, which have been prepared and are available in entirety to the public.¹⁰ Containing 203 reports dating from 2003 to the current year, the briefs cover a wide range of topics, including patient populations, disease states, payer status, quality of care, and quality of care indicators.¹⁰

Publication requirements for research produced from use of HCUP data are very specific. Users of HCUP datasets are required to complete and submit evidence of completing HCUP Data Use Agreement Training.¹¹ According to AHRQ, privacy protections must be verified, including non-disclosure of individual persons, either directly or indirectly; non-disclosure of hospitals; and avoidance of publication of cell sizes less than or equal to 10. Specific instructions for including citations of each database, HCUP tools, and HCUPnet in both abstracts and manuscripts are also offered on the HCUP Web site.¹³

Challenges associated with Healthcare Cost and Utilization Project Nationwide Inpatient Sample

Big data and large datasets are traditionally associated with data input from different users, resulting in a high degree of heterogeneity. Despite the size of the HCUP datasets, there is a high degree of uniformity. The HCUP data is collected by state-level partners, who then submit to HCUP NIS where the data is further organized in standardized fields, making manipulation of the files somewhat less onerous for researchers. The size of the datasets, sampling strategy used to create them, and interpretation of terms posed challenges to the research team.

Data extraction

Extracting data from the central databases is often not a straightforward process. While the field headings are well described both online and in accompanying information sent with the datasets, additional expertise to interpret coding within the fields may be needed. For example, users can quickly note that diagnoses are stored in fields DX1-DX15, interpretation of codes from International Classification of Diseases, 9th revision (ICD-9)¹⁵ and understanding the relevance of the numbering may require a clinician's knowledge. Without the capability to utilize the aforementioned AHRQ provided software, development of reading and extraction software will be required of the research group. Because of the dataset variable inconsistency across datasets from year to year and database redesigns, developing systematic commands is difficult.

The HCUP NIS datasets are designed to protect the privacy of contributing hospitals by hiding and removing data that may reveal hospital identification and location. However, this strategy also creates barriers for researchers to answer and study many important questions. Lacking lower level identifiers also creates difficulties to link HCUP data to other datasets, such as pollution data on the city and state level.

Sampling methods

Because of the sampling methodology, HCUP NIS datasets are ideally suited for determining national level estimates and trends, but are not as useful for obtaining specificity at the level of individual states. In the HCUP NIS datasets, the sampling methodology is designed to create a sample of 20% of inpatient hospital admissions across the United States. 'State' is not included in these datasets as a stratifier, so without additional files analysts cannot accurately generate state-level estimates by using HCUP NIS datasets. Data extracted on the state level are normally unbalanced because of the particular sampling techniques adopted. For instance, when predicting 30-day patient readmission rate using HCUP data, Zhu et al. encountered an unbalanced sample that presented great heterogeneity.¹² In order to deal with sampling issues, they adopted decision tree and logistic models to stratify the sample to subgroups to reduce heterogeneity.¹²

In addition, weights that could be used to calculate state-level estimates are not included in the datasets, though may be found in supplemental files for some states. The supplemental files can be used to enhance the NIS files. In our proposed project, states of particular interest to the research team did not contribute to data to the supplemental state-level files.

Data storage and transmission

Big data requires flexible and easily expandable storage capacity and management solutions.⁵ Storage methods need to be both reliable and also available for easy access while maintaining the minimal defined level of data security.¹⁴ In our research project, each interdisciplinary team member worked on a different aspect of data analysis and interpretation requiring the aforementioned central repository. Challenges arise when researchers are not connected using an internal network. Collaborators across facilities often work on networks of varying capabilities and levels of security, as was the case with our team.

Addressing project challenges

Available technology and software support

Due to the size of the HCUP NIS datasets, a central storage location containing all data from years 2007 to 2011 was not possible given our team's resources. Storage and transmission of large datasets commonly requires institutional support. To address our questions, the data had to be organized in subsets to include only the variables we felt were necessary (Figure 2). These files were placed into encrypted cloud-storage so that researchers, after completing HCUP Data User Training, could access it for their needs. We noted early on that problems could easily arise in naming, altering, and analyzing the files, reinforcing the need for regular communication between the team members on progress with file analysis.

Importance of the inter-disciplinary team

Despite the often robust nature of the data included in large datasets, analysis has typically required the addition of scientists with specialized skillsets to manipulate and derive answers from the massive amounts of raw data. Our team quickly understood the necessity of embracing a combination of skillsets to answer our research questions. At the beginning of the project, our team included domain experts from nursing and atmospheric sciences, who

generated many ideas for possible research efforts and had experience in data management. Our team quickly expanded to include biostatisticians and health policy experts who were able to interpret and apply results in aspects of environmental policy while ensuring that our research methods were methodologically sound. Our team-based approach has been much more efficient in preparing the data for analysis and interpreting its results, while identifying other compelling questions supporting the need for further research throughout the process.

Associated costs

Costs of the project to date are minimal. Researchers may purchase annual HCUP NIS datasets at costs ranging from \$20 to \$500, depending on the year requested and status of the purchase (student vs. others). Data costs for our team were supported by internal funding from the University of Alabama in Huntsville, Office of the Vice President for Research. The smaller size of our research team has enabled us to work efficiently, without additional funding for labor costs. Larger research teams, particularly if the teams include external partners, may require additional funding for activities associated with analysis of the datasets.

Decisions in data use

Dialogue among team members regarding appropriate variables and years for analysis is an ongoing process, emphasizing the need for regular communication. While on first glance it may seem that the years used for longitudinal data analysis (2007 to 2011) might be somewhat dated, our research team chose these years for logistical reasons. Because the project discussion and purchase of the datasets began in 2014, the decision to use 5 years of data was appropriate. The HCUP datasets typically become available for purchase 12 to 24 months after the year of data collection. The sampling methodology for HCUP NIS changed in 2012, from a state-by-state analysis (of participating states) of 20% of inpatient hospital admissions, to a national sample of 20% of hospital admissions. Accommodating this change would have necessitated the modification of datasets to match one another, requiring techniques we did not have at the project's beginning. Five years of environmental data analysis is also the usual time frame needed for the US Environmental Protection Agency to review and offer recommendations regarding changes in levels of fine particulate matter (PM_{2.5}) and other pollutants.

Shifting the analysis paradigm

The availability of large scale datasets enables clinical research to go beyond the traditional cohort, case-control, and clinical trial designs. Data availability creates opportunities for more sophisticated statistical modeling that predict outcomes of interest by using continuous observations. Biostatisticians and health policy analysts, who were included as team members, offered unique perspectives throughout the process of working with the HCUP NIS datasets. The addition of these disciplines has made it possible for the research team to expand efforts, producing epidemiological descriptive statistics and applying explanatory statistical linear modeling techniques that can more fully describe answers that may be available only in large datasets.

Conclusion

Analysis of the HCUP NIS datasets to address our questions has filled a gap identified in current literature, which speaks to the need for analysis of trends in HCUP regions. Our use of HCUP NIS datasets has been helpful in refining our research questions and areas of geographical interest, and helping us better understand our needs for access to highly granular spatial and temporal data to conduct future projects. For example, to be more useful for many researchers, data could be aggregated on a daily basis, instead of monthly. With respect to spatial data, flags designating facility locations according to metropolitan status would be extremely useful for calculating trends. Capital investments by partners who wish to use the HCUP datasets, such as securing computing resources to promote collaboration among multi-disciplinary teams, would increase the efficiency and economy of analysis. Creating secure spaces as data repositories, which would offer access to qualified researchers, would also increase the facility of big data research.

Despite the challenges our team encountered in using the HCUP NIS datasets, we remain convinced that the economic and temporal expenses incurred in using them are a good investment for our area of research. We have a better appreciation of the need for multi-disciplinary interpretation of the data in addressing complex problems which are without straightforward answers, such linking physical processes (such as changes in PM_{2.5}) that have an impact on public health. The challenges that we encountered will inform further efforts as our project evolves, both in continued use of publicly available datasets and collaboration with industry partners.

Acknowledgments

The participation of Robert A Oster was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under award number UL1TR001417.

References

1. Demchenko, Y., Zhao, Z., Grosso, P., Wibisono, A., De Laat, C. Addressing big data challenges for scientific data infrastructure. Cloud Computing Technology and Science (CloudCom), 2012 IEEE 4th International Conference on; IEEE; 2012 Dec. p. 6-617.
2. Ammu N, Irfanuddin M. Big Data Challenges. International Journal of Advanced Trends in Computer Science and Engineering. 2013; 2(1):613–615.
3. Groves P, Kayyali B, Knott D, Van Kuiken S. The 'big data' revolution in healthcare. McKinsey Quarterly. 2013;2.
4. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health Information Science and Systems. 2014; 2(1):1–10. [PubMed: 25825665]
5. Roski J, Bo-Linn GW, Andrews TA. Creating value in health care through big data: opportunities and policy implications. Health Affairs. 2014; 33(7):1115–1122. [PubMed: 25006136]
6. Piai, S., Claps, M. [Accessed June 1, 2016] Bigger data for better healthcare. IDC Health Insight, White Paper Web site. 2013. <http://www.intel.com/content/www/us/en/healthcare-it/solutions/documents/bigger-data-better-healthcare-idc-insights-white-paper.html>
7. Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Affairs. 2014; 33(7):1123–1131. [PubMed: 25006137]

8. Agency for Healthcare Research and Quality, Healthcare Cost and Utilization Project (HCUP). [Accessed August 1, 2016] Overview of the National (Nationwide) Inpatient Sample Web site. 2011. www.hcup-us.ahrq.gov/nisoverview.jsp
9. HCUP Nationwide Inpatient Sample (NIS). Healthcare Cost and Utilization Project (HCUP). Agency for Healthcare Research and Quality; Rockville, MD: 2011. www.hcup-us.ahrq.gov/nisoverview.jsp [Accessed August 1, 2016]
10. HCUP Nationwide Inpatient Sample (NIS). Healthcare Cost and Utilization Project (HCUP). [Accessed August 5, 2016] Reports Web site. <https://www.hcup-us.ahrq.gov/reports/pubsearch/pubsearch.jsp>
11. HCUP Nationwide Inpatient Sample (NIS). Healthcare Cost and Utilization Project (HCUP). [Accessed June 1, 2016] HCUP User Data Training Web site. https://www.hcup-us.ahrq.gov/tech_assist/dua.jsp
12. Zhu K, Lou Z, Zhou J, Ballester N, Kong N, Parikh P. Predicting 30-day Hospital Readmission with Publicly Available Administrative Database. *Methods of information in medicine*. 2015; 54(6):560–567. [PubMed: 26548400]
13. Healthcare Cost and Utilization Project (HCUP). [Accessed May 1, 2016] Citations for HCUP Databases and Tools Web site. https://www.hcup-us.ahrq.gov/tech_assist/citations.jsp
14. Chen M, Mao S, Liu Y. Big data: a survey. *Mobile Networks and Applications*. 2014; 19(2):171–209.
15. Medicode (Firm). ICD-9-CM: International Classification of Diseases, 9th Revision, Clinical Modification. Salt Lake City, Utah: Medicode; 1996.
16. Biel, TJ. PyHCUP 0.1.6.4. 2015. <https://pypi.python.org/pypi/PyHCUP/0.1.6.4>

Key Points

- Large datasets are valuable resources for healthcare research and can be used to provide robust answers for clinical questions
- Preparation of large datasets for research purposes can be complex and time-consuming tasks
- The varied skillsets of a multi-disciplinary team can be useful in querying large datasets for answers to research questions
- Ongoing communication and flexibility in strategic management can help the research team to get the greatest value from using large datasets

Figure 2.
Simple depiction of data processing and flow, with approximate file sizes, as it moves throughout an integrated science team